International Journal of Approximate Reasoning 51 (2010) 224–239

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

A Bayesian network model for spatial event surveillance

Xia Jiang [a,*], Daniel B. Neill [b], Gregory F. Cooper [a]

[a] Department of Biomedical Informatics, University of Pittsburgh, Parkvale Building M183, 200 Meyran Ave., Pittsburgh, PA 15260, United States
[b] H.J. Heinz III College, School of Public Policy and Management, Carnegie Mellon University, Pittsburgh, PA 15213, United States

ARTICLE INFO

Article history:
Received 19 March 2008
Received in revised form 8 August 2008
Accepted 5 January 2009
Available online 13 January 2009

Keywords:
Bayesian network
Spatial cluster detection
Event surveillance
Spatial event surveillance
Outbreak detection

ABSTRACT

Methods for spatial cluster detection attempt to locate spatial subregions of some larger region where the count of some occurrences is higher than expected. Event surveillance consists of monitoring a region in order to detect emerging patterns that are indicative of some event of interest. In spatial event surveillance, we search for emerging patterns in spatial subregions.

A well-known method for spatial cluster detection is Kulldorff's [M. Kulldorff, A spatial scan statistic, Communications in Statistics: Theory and Methods 26 (6) (1997)] spatial scan statistic, which directly analyzes the counts of occurrences in the subregions. Neill et al. [D.B. Neill, A.W. Moore, G.F. Cooper, A Bayesian spatial scan statistic, Advances in Neural Information Processing Systems (NIPS) 18 (2005)] developed a Bayesian spatial scan statistic called BSS, which also directly analyzes the counts.

We developed a new Bayesian-network-based spatial scan statistic, called BNetScan, which models the relationships among the events of interest and the observable events using a Bayesian network. BNetScan is an entity-based Bayesian network that models the underlying state and observable variables for each individual in a population.

We compared the performance of BNetScan to Kulldorff's spatial scan statistic and BSS using simulated outbreaks of influenza and cryptosporidiosis injected into real Emergency Department data from Allegheny County, Pennsylvania. It is an open question whether we can obtain acceptable results using a Bayesian network if the probability distributions in the network do not closely reflect reality, and thus, we examined the robustness of BNetScan relative to the probability distributions used to generate the data in the experiments. Our results indicate that BNetScan outperforms the other methods and its performance is robust relative to the probability distribution that is used to generate the data.

1. Introduction

Methods for spatial cluster detection attempt to locate spatial subregions of some larger region where the count of some occurrences is higher than expected. As an example, we may look for clusters of certain kinds of trees in a forest; other applications of spatial cluster detection include mining astronomical data, medical imaging, and military surveillance [1]. Event surveillance consists of monitoring a region in order to detect emerging patterns that are indicative of some event of interest. As examples, we may look for emerging patterns that are indicative of a disaster that is in its early stages of development. Examples of such disasters include hurricanes, terrorist attacks, and outbreaks of diseases. In spatial event surveillance, we search for emerging patterns in spatial subregions. Spatial cluster detection is one statistical technique used for spatial event surveillance.

* Corresponding author.
E-mail addresses: xjiang@cbmi.pitt.edu (X. Jiang), neill@cs.cmu.edu (D.B. Neill), gfc@pitt.edu (G.F. Cooper).

0888-613X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2009.01.001

A well-known method for spatial cluster detection is Kulldorff's [1] spatial scan statistic, which has been implemented as freeware in the SaTScan™ software package [2]. This method directly analyzes the counts of occurrences in the various subregions. Neill et al. [3] developed a Bayesian spatial scan statistic called BSS, which also directly analyzes the counts. When the clusters of interest are clusters of the events we observe, then directly analyzing the counts is likely to be best. However, in event surveillance, and disease-outbreak detection in particular, we ordinarily observe events that are *related* to the occurrences of interest. As an example, when we are interested in whether there is an outbreak of a certain disease, we observe individuals with various symptoms of the disease rather than the disease infections themselves. Instead of using a summary statistic, we develop a Bayesian-network-based spatial scan statistic, called BNetScan, which models the relationships among the events of interest and the observable events using a Bayesian network. This network is then used to determine the posterior probability of each subregion containing a cluster. A strength of this method is that it can model multiple causes of the clusters. For example, in disease-outbreak detection it can model any number of possible disease outbreaks using a single Bayesian network.

An important question is whether we can obtain acceptable results using a Bayesian network if the probability distributions in the network, which are often obtained from limited data and/or subjective judgment, do not closely reflect reality. We describe the results of experiments that test the robustness of BNetScan relative to changes in the probability distributions used to generate the experimental data. In particular, we examine the ability of BNetScan to detect outbreaks of influenza and cryptosporidiosis using simulated cases injected into real Emergency Department data from Allegheny County, Pennsylvania. As a point of comparison, we included SaTScan™ and BSS in the study, and we evaluated these systems both when they took advantage of the probabilistic information in BNetScan (i.e., which chief complaints are most commonly associated with each outbreak type) and when they did not. In this way, we could learn whether it is simply the use of probabilistic information that accounts for a certain performance level, or whether that performance may be due to the use of a Bayesian network model.

In the remainder of this section, we provide background on spatial cluster detection. Section 2 describes the two methods, SaTScan™ and BSS, which we compare to BNetScan. In Section 3 we develop our new Bayesian-network-based spatial scan statistic method, BNetScan. Section 4 presents our experiments and their results. Finally, Section 5 provides a discussion of the results.

1.1. Spatial cluster detection

In spatial cluster detection, the goal is to identify the location, shape, and size of possible clusters, and to determine how likely it is that a cluster is due to the event with which we are concerned (e.g., a disease outbreak) versus how likely it is that the cluster is merely a chance occurrence. When doing spatial cluster detection, we first enumerate the subregions of some geographical region G. For example, Kulldorff [1] places a circular window over the region and lets the center of the circle move over the region. For each center, the radius of the circle is varied. Neill et al. [3] represent the entire region by a grid and search over rectangular subregions of the grid. Inherent in these methods is that we assume that the entire region G is composed of cells. For example, if we cover G with an $m \times n$ grid, each grid element is a cell. A subregion of G is the union of any number of cells. Since the number of such subregions is exponentially large, Neill et al. [3] search only over subregions which are axis-aligned rectangles. We take this same approach in this paper.

Classical methods for spatial cluster detection detect clusters based solely on counts of occurrences in the various subregions. Kulldorff [1,4]) developed the well-known frequentist method called the spatial scan statistic. The scan statistic was first proposed by Naus [5] as a solution to the multiple hypothesis testing problem. Scan statistics have been used to find clusters of chronic diseases such as breast cancer [6] and leukemia [7]. They have also been used to detect clusters of work-related hazards [8] and West Nile virus [9]. As mentioned earlier, Kulldorff implemented the spatial scan statistic in his SaTScan™ software [2]. Jung et al. [10] develop a version of the spatial scan statistic that considers multinomial variables whose values are ordinal. Neill et al. [3] developed a Bayesian spatial scan statistic, and Neill et al. [11] developed a multivariate version of the Bayesian spatial scan statistic.

The methods discussed so far are non-temporal. That is, if we are looking for patterns that emerge with time, they look only at data from the most recent time period. A temporal method would detect a cluster based not only on data from the most recent time period, but also on data from previous time periods. Kulldorff et al. [12] developed a temporal version of the spatial scan statistic that looks at three-dimensional cylinders, while Neill et al. [13] developed a temporal version of the spatial scan statistic that can detect emerging clusters with incidence rates that increase over time.

In general, spatial cluster detection does not necessarily entail the notion of time. In some applications we many want to determine whether there is currently a cluster without concern for whether the counts are changing with time. For example, we may want to investigate whether there is a cluster of a particular type of star in space. Next we discuss spatial event surveillance, which does concern counts changing with time.

1.2. Spatial event surveillance

In event surveillance we try to detect emerging patterns that are indicative of some forthcoming event or an event that is in its early stages. A classical example is disease-outbreak detection. A disease outbreak detection system monitors a region each day to see if some pattern has emerged that is indicative of a disease outbreak. For example, a *Cryptosporidium* infection

can cause diarrhea. So the count of over-the-counter (OTC) sales of antidiarrheal medications typically increases during a *Cryptosporidium* outbreak. We can therefore analyze emerging patterns of antidiarrheal medications to try to learn whether a *Cryptosporidium* outbreak has started. In spatial event surveillance we look for emerging patterns in spatial subregions. In this way we may detect an emerging pattern earlier than if we analyzed a region globally. For example, in the case of a *Cryptosporidium* outbreak, the outbreak may start near the location of a contaminated drinking water supply. We may detect an emerging pattern in a small subregion containing the water supply before we could detect an emerging pattern in the entire city or county.

2. Previous methods for spatial cluster detection

Next we describe the two spatial cluster detection statistics which were compared to BNetScan in our experiments.

2.1. Kulldorff's spatial scan statistic

Kulldorff [1] assumes the entire region G is composed of cells c_i. For example, if we cover G with an $m \times n$ grid, each grid element is a cell. A subregion s of G is the union of any number of cells. We are interested in whether some subregion s of G contains a cluster. The null hypothesis H_0 is that there are no clusters, and the alternative hypothesis H_s is that subregion s contains a cluster. Kulldorff [1] developed two different scan statistic models, the Bernoulli model and the Poisson model. The statements of the hypotheses are different in the two models. We discuss each in turn.

In the Bernoulli model, every cell $c_i \subseteq G$ contains a discrete number of entities. Each entity either does or does not have some property X. For example, the entities could be people, and the property could be that a person visited the Emergency Department with a cough. We would then be interested if there is a cluster of such entities in some subregion of a particular city or county, for instance. As another example, the entities could be stars, and the property could be that the star is of a particular type. We may be interested in whether there is a cluster of this particular type in some subregion of space.

The statements of the hypotheses for the Bernoulli model are as follows:

H_0: For an entity in the entire region G, the probability of the entity having property X is q. The event of any one entity having the property is independent of another entity having it.
H_s: For an entity in subregion s, the probability of the entity having property X is p. For an entity in subregion $G - s$, the probability of the entity having property X is q, where $p > q$. The event of any one entity having the property is independent of another entity having it.

We define the following variables:

B: the total number of entities in G,
$B_{in}^{(s)}$: the total number of entities in s,
$B_{out}^{(s)}$: the total number of entities in $G - s$,
C: the total number of entities in G with property X,
$C_{in}^{(s)}$: the total number of entities in s with property X,
$C_{out}^{(s)}$: the total number of entities in $G - s$ with property X.

Kulldorff [1] shows that the most significant spatial cluster is the subregion s that maximizes the following likelihood ratio statistic:

$$L(s) = \frac{\left(\frac{C_{in}^{(s)}}{B_{in}^{(s)}}\right)^{C_{in}^{(s)}} \left(1 - \frac{C_{in}^{(s)}}{B_{in}^{(s)}}\right)^{B_{in} - C_{in}^{(s)}} \left(\frac{C_{out}^{(s)}}{B_{out}^{(s)}}\right)^{C_{out}^{(s)}} \left(1 - \frac{C_{out}^{(s)}}{B_{out}^{(s)}}\right)^{B_{out} - C_{out}^{(s)}}}{\left(\frac{C}{B}\right)^{C} \left(1 - \frac{C}{B}\right)^{B-C}}$$

if

$$\frac{C_{in}^{(s)}}{B_{in}^{(s)}} > \frac{C_{out}^{(s)}}{B_{out}^{(s)}},$$

otherwise

$$L(s) = 1.$$

If subregion s^* has the highest value of the test statistic among all the subregions being tested, we only know that s^* is the most likely subregion to contain a cluster. We determine the statistical significance of this subregion using Monte Carlo simulation. The technique was originally proposed in [14], and it was first used in the context of a scan statistic in [15]. In this technique, we obtain a large number N of replications of the data set, each of which is generated under the null hypothesis. The p-value of H_s is then equal to

$$\frac{N_{beat}+1}{N+1},$$

where N_{beat} is the number of replications in which the subregion with the highest value of the test statistic has a higher value than $L(s^*)$. For example, with 999 such replications, the test is significant at the 0.05 level if $L(s^*)$ is among the 50 highest values obtained from the replications.

In the Poisson spatial scan model, every cell $c_i \subseteq G$ contains a variable number of points (entities), which we count. For example, the count may be the number of entities with some property. It is assumed that counts are being generated according to an inhomogeneous Poisson process. The statements of the hypotheses are as follows:

H_0: For every cell c_i, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(qb_i),$$

where b_i is a baseline count associated with cell c_i. Kulldorff [1] uses the underlying population of each cell as its baseline, while Neill et al. [3] use an expected count estimated from time series analysis of historical data.

H_s: For cells $c_i \subset s$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(pb_i)$$

and for cells $c_i \not\subseteq s$, the counts are distributed as follows:

$$C_i \sim \text{PoissonDist}(qb_i),$$

where $p > q$.

Define the following variables:

$$B = \sum_{c_i} b_i,$$
$$B_{in}^{(s)} = \sum_{c_i \subset s} b_i,$$
$$B_{out}^{(s)} = \sum_{c_i \not\subseteq s} b_i,$$
$$C = \sum_{c_i} C_i,$$
$$C_{in}^{(s)} = \sum_{c_i \subset s} C_i,$$
$$C_{out}^{(s)} = \sum_{c_i \not\subseteq s} C_i.$$

Kulldorff [1] shows that the most significant subregion s is the one that maximizes the following Poisson spatial scan statistic:

$$F(s) = \frac{\left(\frac{C_{in}^{(s)}}{B_{in}^{(s)}}\right)^{C_{in}^{(s)}} \left(\frac{C_{out}^{(s)}}{B_{out}^{(s)}}\right)^{C_{out}^{(s)}}}{\left(\frac{C}{B}\right)^{C}} \tag{1}$$

if

$$\frac{C_{in}^{(s)}}{B_{in}^{(s)}} > \frac{C_{out}^{(s)}}{B_{out}^{(s)}},$$

otherwise

$$F(s) = 1.$$

As is the case for the Bernoulli spatial scan statistic, we determine the statistical significance of our finding by using Monte Carlo simulation.

2.2. The Bayesian spatial scan statistic

The Bayesian spatial scan statistic (BSS) [3] is a Bayesian method for spatial cluster detection, which allows us to incorporate prior information and to calculate the posterior probability of each spatial subregion. Specifically, this method places a prior gamma probability distribution on the values of q, and a mixture of gamma distributions on p, all of which have means greater than the mean of the gamma distribution on q. It then places Poisson distributions on the counts conditional on these values of p and q. Furthermore, it places a prior probability on the event that there is a cluster in each subregion.

Finally, it computes the posterior probability of an outbreak in each subregion based on the data. Since it determines posterior probabilities, there is no need to create replications of the data set and to determine significance, which greatly decreases its time complexity relative to the frequentist spatial scan statistic.

3. The Bayesian network spatial scan statistic (BNetScan)

When the clusters of interest are clusters of the events that we directly observe, then directly analyzing the counts is likely to be best. However, in event surveillance, and disease-outbreak detection in particular, we ordinarily observe events that are *related* to the events of interest. As an example, an individual with *Cryptosporidium* infection may demonstrate certain Emergency Department (ED) symptoms or purchase certain over-the-counter (OTC) medications. Such an individual might visit the ED with diarrhea or purchase antidiarrheal medication. We observe the ED symptoms and medication purchases, but we do not observe the *Cryptosporidium*infection, which is the occurrence of interest. We could try to detect a *Cryptosporidium* outbreak by using the spatial scan statistic to analyze the counts of ED visits or OTC sales. However, another approach is to model the probabilistic relationships among a *Cryptosporidium* outbreak, individuals being infected with *Cryptosporidium*, individuals presenting in the ED with indicators of *Cryptosporidium*, and individuals purchasing OTC medications related to a *Cryptosporidium* infection. Our data then consists of the values of the observed variables for all individuals in the population, and we compute the posterior probability of an outbreak given these data. This is a generative, multi-entity, model-based approach to spatial cluster detection, since we model the underlying state and observed variables for each entity or individual in the population. A strength of this method is that it can model multiple causes of the clusters. For instance, it could model a *Cryptosporidium* outbreak, an influenza (influenza) outbreak, and other disease outbreaks using a single Bayesian network.

We first describe a very simple version of the method to illustrate the basic concepts. Then we present the general method, which we call the Bayesian network spatial scan statistic (BNetScan). We then describe how we used that method to develop a specific disease-outbreak detection system. Section 4 presents an evaluation of the system.

3.1. A simple formulation of the model

Fig. 1 shows a Bayesian network representing a simple multi-entity spatial cluster detection system. As a concrete example, consider a network designed to detect influenza outbreaks. The variable E has value "yes" if there is currently an outbreak of influenza and value "no" otherwise. The variable S has value s_j if there is an outbreak in subregion s_j and has value "none" if there is no outbreak in any subregion. Recall that a subregion is any subset of cells. In our applications, we cover the entire region with an $m \times n$ grid, where each grid element is a cell, and we only consider subregions which are rectangles. We must specify the following spatial prior probability: $P(S = s_j | E = yes) = a_j$, where $\sum_j a_j = 1$. The values of a_j could all be the same, or they could be based, for example, on the size and regularity of the subregions.

There is a variable I_r for each individual r in region G. The possible values of I_r are our possible observations o_k for each individual. In this example, they are the chief complaints with which the individual might present in the Emergency Department, where one value is "noED", which means the individual did not visit the Emergency Department. We see from Fig. 1 that if there is an outbreak in subregion s_j, individuals in s_j have probability p_k of having chief complaint o_k, while individuals not in s_j have probability q_k of having chief complaint o_k. If there is no outbreak in any subregion, all individuals have

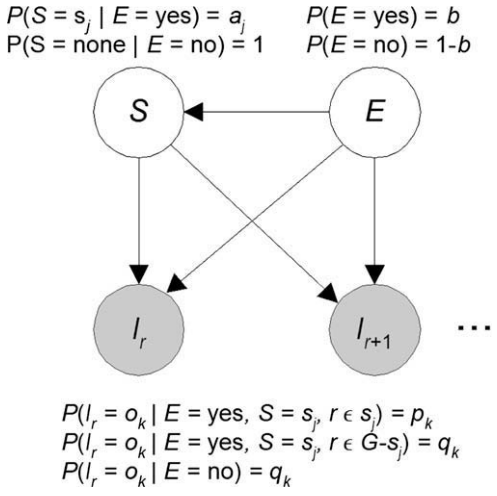

Fig. 1. A Bayesian network for a multi-entity spatial cluster detection system. The shaded variables are the ones we observe.

probability q_k of having chief complaint o_k. Note that for each assignment of values to E and S, the Bayesian network in Fig. 1 is a naive Bayesian network.

Our *Data* consists of the values of I_r for all individuals r in G. These nodes are shaded to indicate that they represent measured variables. Since there could be thousands, or even millions, of individuals in G, we would not explicitly construct the Bayesian network in Fig. 1, and instantiate I_r for all r. Rather we can compute the likelihood of the data as follows:

$$P(Data|S = s_j) = \prod_k (p_k)^{C_k^{(j)}} (q_k)^{N_k - C_k^{(j)}}$$
$$P(Data|E = yes) = \sum_j P(Data|S = s_j) a_j$$
$$P(Data|E = no) = \prod_k (q_k)^{N_k},$$

where $C_k^{(j)}$ is the number of individuals in s_j with the kth chief complaint, and N_k is the number of individuals in G with the kth chief complaint. We then use Bayes' rule to compute posterior probabilities as follows:

$$P(E = yes|Data) = \frac{P(Data|E=yes)P(E=yes)}{P(Data|E=yes)P(E=yes)+P(Data|E=no)P(E=no)}$$
$$P(S = s_j|Data) = \frac{P(Data|S=s_j)P(S=s_j)}{\sum_i P(Data|S=s_i)P(S=s_i)},$$

where one value of s_i is "none".

3.2. The general formulation of the model

Fig. 2 depicts the general structure of a multi-entity spatial cluster detection Bayesian network. There is a node E whose values are the possible events that could be occurring. There is a node S whose values are the possible subregions in which an event could be occurring. There are one or more observable nodes for each individual r in the region. We show one node labeled I_r, but in general I_r represents an entire Bayesian subnetwork for individual r. There can be any number of other nodes (represented abstractly in Fig. 2 as *Intermediate Nodes*) that convey the relationships among the observable nodes and S and E. In the next subsection, we describe a model that concretely illustrates intermediate nodes.

3.3. An example of the general model

This section describes a multi-entity spatial cluster detection system that is based on the PANDA-CDCA Bayesian network. PANDA-CDCA [16] is a non-spatial disease-outbreak detection system. It monitors an entire region globally in an effort to detect an outbreak of disease. PANDA-CDCA models the CDC Category A diseases, namely, anthrax, plague, smallpox, tularemia, botulism and hemorrhagic fever, and also certain diseases that may be confused with them, namely influenza, *Cryptosporidium*, and hepatitis. It includes variables for each individual in a population.

Fig. 3 shows the Bayesian network in PANDA-CDCA modified to handle spatial cluster detection. We call this Bayesian network BNetScan-PC. The probability distributions in the network are relative to the most recent time period, which consists of the past 24 h. The value of the variable E is the type of outbreak that is occurring if there is an outbreak. It is assumed that at most one type of outbreak will be occurring at any given time. There are 13 possible values of E. One of the values is "none", which means there is no outbreak. The remaining 12 values are the outbreak diseases, with plague, anthrax, and

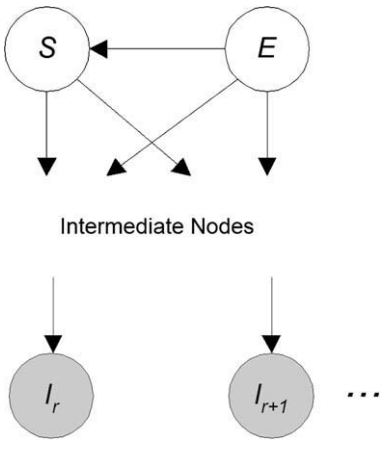

Fig. 2. The general structure of a multi-entity spatial cluster detection Bayesian network.

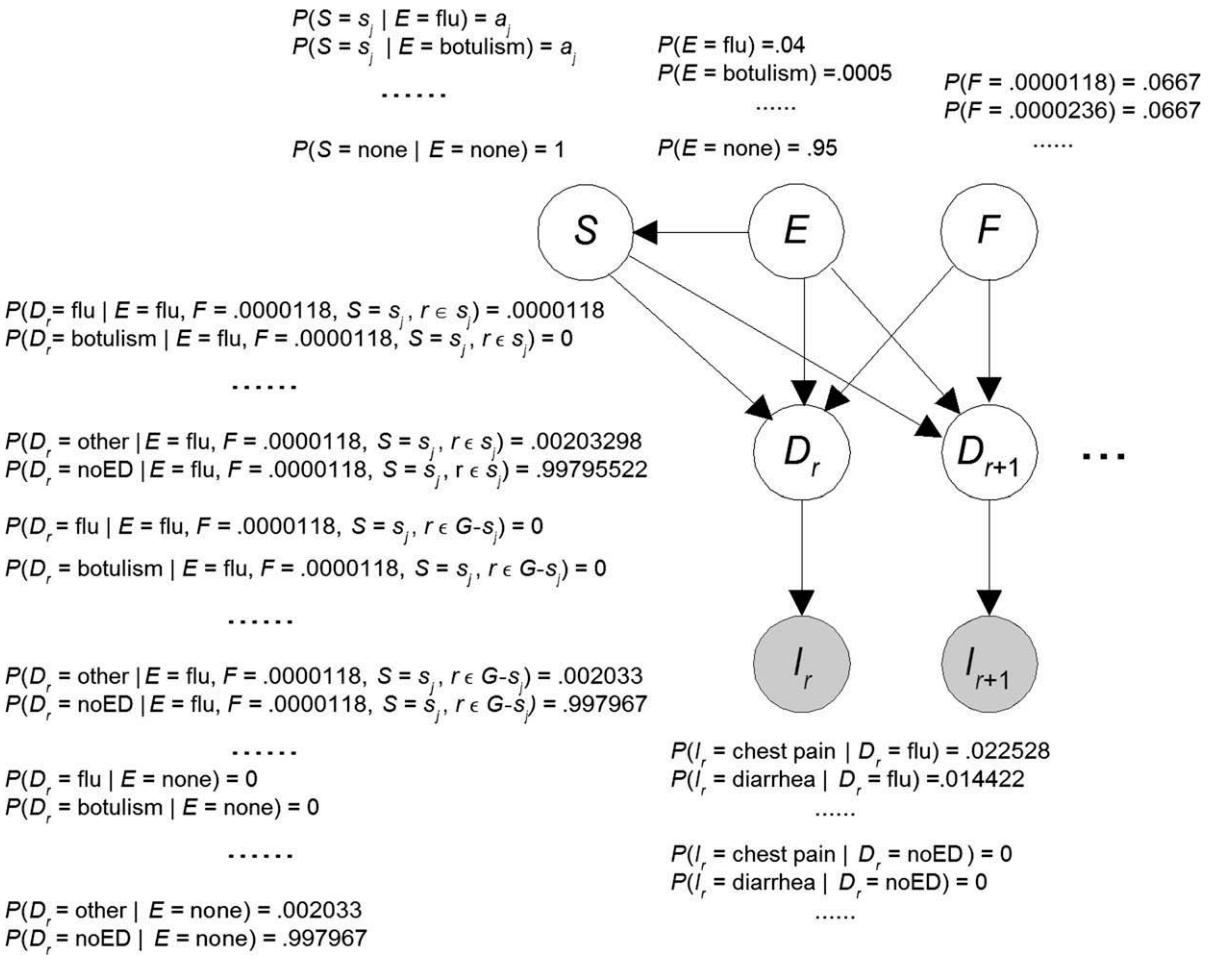

Fig. 3. The Bayesian network in PANDA-CDCA modified to handle spatial cluster detection.

hemorrhagic fever each occurring twice, once in an early stage and once in a late stage (since the early-stage and late-stage symptoms of these diseases are very different).

The value of the variable S is the spatial subregion s_j in which the outbreak is occurring, if there is an outbreak. Its value is "none" if there is no outbreak.

There are variables D_r and I_r for each individual r in region G. The value of the variable D_r is the disease state of the rth individual if the individual visited the ED and is "noED" otherwise. There are 14 possible disease states including every outbreak disease, asthma, and other. Variable D_r is an unmeasured (hidden) variable in the sense that we can only infer its value under uncertainty. The value "other" for D_r means the individual visited the ED with something other than an outbreak disease. For example, the individual may have visited the ED with a broken arm. The value of the variable I_r is the rth individual's chief complaint in the ED if the individual visited the ED, and is "noED" if the individual did not visit the ED.

Variable I_r has 54 possible values. Fifty-two of the values represent specific chief complaints (e.g., "cough"). Another value of I_r is "other", which denotes all other chief complaints. The final value of I_r is "noED", which indicates individuals who did not visit the ED recently.

The value of the variable F represents the extent of the outbreak, if one is occurring. For example,

$$P(D_r = \textit{influenza} | E = \textit{influenza}, F = 0.0000118, S = s_j, r \in s_j) = 0.0000118,$$
$$P(D_r = \textit{influenza} | E = \textit{influenza}, F = 0.0000236, S = s_j, r \in s_j) = 0.0000236.$$

So if the extent of an influenza outbreak in subregion s_j is $F = 0.0000118$, an individual in s_j has probability 0.0000118 of arriving in the ED with influenza, whereas if the extent is $F = 0.0000236$, an individual in s_j has probability 0.0000236 of arriving in the ED with influenza.

Our *Data* consists of the values of I_r for all individuals r in G. As in Section 3.1, we need not actually construct the Bayesian network in Fig. 3, and instantiate I_r for all r. Rather we can compute the likelihood of the data in the way we show next. Owing to the conditional independencies entailed by the network, we have that

$$P(Data|E,F,S=s_j)=\prod_{r\in s_j}\sum_{D_r}P(I_r|D_r)P(D_r|E,F,S=s_j,r)\prod_{r\in G-s_j}\sum_{D_r}P(I_r|D_r)P(D_r|E,F,S=s_j,r).$$

Let D be a random variable whose possible values are the disease states, and I be a random variable whose possible values are the chief complaints. That is, D and I have the same spaces as D_r and I_r, respectively, but they do not refer to any particular individual. For each of the 54 possible values of I, we need only compute two values of $\sum_D P(I|D)P(D|E,F,S=s_j,r)$, one for an individual in s_j and one for an individual outside s_j. To that end, let

$$p_k=\sum_D P(I=o_k|D)P(D|E,F,S=s_j,r\in s_j),$$
$$q_k=\sum_D P(I=o_k|D)P(D|E,F,S=s_j,r\in G-s_j),$$

where o_k is the kth chief complaint. Note that p_k and q_k do not depend on s_j, but they do depend on E and F, which we leave implicit for notational simplicity. We then have

$$P(Data|E,F,S=s_j)=\prod_{k=1}^{54}(p_k)^{C_k^{(j)}}(q_k)^{N_k-C_k^{(j)}},$$

where $C_k^{(j)}$ is the number of individuals in s_j with the kth chief complaint, and N_k is the number of individuals in G with the kth chief complaint. We have next that

$$P(Data|S=s_j)=\sum_{E\neq none}\sum_F P(Data|E,F,S=s_j)P(F)P(E|S=s_j)$$

and

$$P(Data|E)=\sum_{s_j}\sum_F P(Data|E,F,S=s_j)P(F)P(S=s_j|E).$$

Once we have these probabilities, we proceed to apply Bayes' rule in the same way as shown at the end of Section 3.1.

4. Experiments

This section describes experiments that compare the detection power of SaTScan™, the Bayesian spatial scan statistic (BSS) method, and BNetScan-PC method. An important question is whether we can obtain good results using a Bayesian network model even if its conditional probability distributions are significantly different from the true generative distributions in the domain being modeled. We tested the robustness of the Bayesian network approach regarding the probability distributions used to generate the outbreak data. In these experiments, we used the SaTScan™ software package, which is available for free at http://www.satscan.org/, in order to implement the spatial scan statistic. We used an implementation of BSS that was developed by Daniel Neill. We used our own implementation of BNetScan-PC. SaTScan™ searched over circular subregions, while BSS and BNetScan-PC searched over rectangular subregions.

4.1. Method

We modeled Allegheny County, Pennsylvania, which covers 730 square miles, using a 16×16 grid. Each grid element is one cell. We considered a zip code to be entirely within a cell if the zip code's centroid was in the cell. We simulated both influenza and *Cryptosporidium* outbreaks in rectangular subregions of the county. Next we discuss the outbreaks. This discussion pertains to both the influenza and *Cryptosporidium* outbreaks.

First, we describe properties of the outbreaks.

(1) *Outbreak Severity*: For each cell, we determined the mean and standard deviation σ_{cell} of the number of ED visits during a 1-year background period when it was assumed no outbreak was occurring. The period chosen was the entire calendar year 2004. The data for each outbreak consisted of new injected ED visits plus the ED data in the background period. So our data is semi-synthetic because the background data is real and the overlaid outbreak data is synthetic. We modeled moderate outbreaks by making the average daily number of injected ED visits in each cell equal to $2\sigma_{cell}$. For an outbreak with duration *dur*, we then let $tot_{cell}=2\sigma_{cell}\times dur$ be the total number of injected ED visits in a given cell during the entire outbreak.

(2) *Daily increase*: We assumed half of the injected ED visits occurred during the first half of the outbreak, and that Δ of them occurred on day one of the outbreak, 2Δ occurred on day two, and so on. Therefore, to determine the value of Δ we solved

$$\Delta+2\Delta+\cdots+\frac{dur}{2}\Delta=\frac{tot_{cell}}{2}.$$

(3) We then injected Δ new outbreak cases into the background data on day one of the outbreak, 2Δ on day two, and so on.

(4) *Chief Complaints*: To determine the chief complaint of each injected case, we generated the chief complaint according to probability distribution P of the chief complaints given the outbreak disease (influenza or *Cryptosporidium*). Recall BNetScan-PC contains a probability distribution P' of the chief complaints given the outbreak disease. To test the robustness of BNetScan-PC, we let P vary significantly from P'. Later, we describe the variation and how many different probability distributions we used. Presently, we show how each distribution was created.To obtain a conditional probability distribution P, we let the conditional probabilities of the chief complaints in BNetScan-PC be the means of Dirichlet distributions. For example, suppose we are simulating an influenza outbreak; if $p_1, p_2, \ldots$, and p_{54} are the conditional probabilities in BNetScan-PC of each of the chief complaints given influenza, and N is our subjective prior sample size, we set

$$a_i = p_i N$$

to obtain parameters for a Dirichlet distribution. Once we have such a Dirichlet distribution, we randomly generate a probability distribution P according to this distribution. This P is then used for one entire set of experiments. As N increases, the likelihood of P being similar to the conditional probability distribution in BNetScan-PC increases.

(5) *Outbreak subregions*: We considered outbreaks that occur in rectangles that are 2 cells by 1 cell, 2 cells by 2 cells, and 3 cells by 2 cells. The 2 by 1 rectangles and 3 by 2 rectangles could go either north–south or east–west. Not all cells had sufficient background ED visits to qualify for containing an outbreak. We eliminated all those cells that had $\sigma_{cell} < 1$. We generated each 2 by 1 rectangle, for example, by randomly choosing two contiguous cells. If either cell had $\sigma_{cell} < 1$, we would randomly generate a new rectangle. We did this until we obtained a 2 by 1 rectangle that did not have $\sigma_{cell} < 1$ in either cell. In this manner, we determined four subregions of each type.

For each type of outbreak (influenza and *Cryptosporidium*), we developed the number of outbreaks described by the table that follows. Recall that N is our subjective prior sample size. By $N = \infty$, we mean that we used the probability distributions in BNetScan-PC. The table also shows the average value of the Kullback–Leibler distances of each generated probability distribution relative to the probability distribution in BNetScan-PC.

N	# Distributions generated	# Outbreaks per distribution	Average KL-Dist. (influenza)	Average KL-Dist. (*Cryptosporidium*)
∞	1	240	0	0
30	10	60	0.24	1.85
5	10	60	1.54	5.86
1	10	60	10.97	22.88

For $N = \infty$, the properties of the 240 outbreaks were determined as follows:

Variable	Values	# Occurrences of each value	Total # of occurrences
Duration	30, 40, 50, 60	60	240
Month	1–12	20	240
Day	1–30	8	240
Subregion	4 of type 2 by 1, 4 of type 2, 4 of type 3 by 2	20	240

The *Month* and *Day* variables determined the starting date for the outbreak. For example, if *Month* = 3 and *Day* = 5, we injected the outbreak data starting on March 5th of the 1-year background period. For each variable, a list of the 240 occurrences was created. To develop an outbreak, an item was sampled at random from each list. The sampled items were removed from the lists before the next outbreak was developed.

For $N = 1$, 5, and 30, the properties of the *60* outbreaks for each probability distribution were determined as follows:

Variable	Values	# Occurrences of each value	Total # of occurrences
Duration	30, 40, 50, 60	15	60
Month	1–12	5	60
Day	1–30	2	60
Subregion	4 of type 2 by 1, 4 of type 2 by 2, 4 of type 3 by 2	5	60

4.2. Results

We evaluated five methods, namely BNetScan, two ways of using BSS, and two ways of using SaTScan™. The two ways we used the latter two systems are as follows. The first way looks for a cluster of individuals presenting in the ED with one of the chief complaints that the outbreak disease could cause according to the probability distribution in BNetScan-PC. For example, in the case of influenza, if an individual presented in the ED with any one of the chief complaints that could be caused by influenza, one is added to the count. There are 12 such chief complaints for *Cryptosporidium* and 20 such chief complaints for influenza. In the second way, we used the probability distribution in BNetScan-PC to determine the three chief complaints which are the best indicator of the outbreak disease. The systems then only look for clusters of individuals presenting with one of these chief complaints. Specifically, in the case of influenza for example, we first assigned a score of 0 to all chief complaints *CC* such that $P(CC—influenza) < 0.002$. In this way we did not include chief complaints that were very unlikely given influenza. Each remaining chief complaint *CC* was assigned a score as follows:

$$score(CC) = \frac{P(CC|D_r = influenza)}{P(CC|D_r = other)}.$$

Recall that the value "other" means the individual visited the ED with something other than an outbreak disease. We then ranked the chief complaints by their scores, and chose the top three chief complaints. In this way, these systems are able to take advantage of the assessed probability distributions in BNetScan-PC despite only being able to incorporate univariate count data. The following tables summarize our results for the top three chief complaints for each outbreak disease:

CC	$P(CC\|D_r = \text{influenza})$	$P(CC\|D_r = \text{other})$	$P(CC\|D_r = \text{influenza}) \div P(CC\|D_r = \text{other})$
Cough	0.3356	0.0248	13.53
Fever/chills	0.4122	0.0322	12.80
Myalgia	0.0095	0.0013	7.31

CC	$P(CC\|D_r = \text{Crypto.})$	$P(CC\|D_r = \text{other})$	$P(CC\|D_r = \text{Crypto.}) \div P(CC\|D_r = \text{other})$
Bloody stools	0.03	0.00005	600
Sweats	0.1375	0.0003	458.33
Diarrhea	0.2643	0.0072	36.71

This table summarizes the inputs to the methods:

Method	Input
BNetScan-PC	The chief complaint of every individual who visited the ED
BSS Method 1	Count of individuals presenting with any chief complaint caused by injected outbreak disease
BSS Method 2	Count of individuals presenting with one of the top three chief complaints
SaTScan™ Method 1	Count of individuals presenting with any chief complaint caused by injected outbreak disease
SaTScan™ Method 2	Count of individuals presenting with one of the top three chief complaints

4.2.1. Outbreak detection

We used AMOC curves [17] to evaluate the ability of the methods to detect the outbreaks. In such curves, the false alarm rate, which is the number of false alarms per year, is plotted on the *x*-axis and the mean number of days to detection is plotted on the *y*-axis. If an outbreak was detected on the first day of the outbreak, we considered the days to detection to be 0; if it was detected on the second day of the outbreak, we considered days to detection to be 1, and so on. If an outbreak was not detected by the mid-point of the outbreak, we set the days to detection to the mid-point. We considered the mid-point of an outbreak to be the last day at which outbreak detection would be useful.

We developed separate AMOC curves for $N = \infty$, $N = 1$, $N = 5$, and $N = 30$. Recall that for the latter three values of N, we generated 10 probability distributions, and developed 60 outbreaks using each distribution. To obtain a *y*-value on the AMOC curve, we took the mean days to detection over all 600 outbreaks.

To obtain points for the AMOC curves, we proceeded as follows. Suppose we were producing an AMOC curve showing how well BNetScan-PC detected whether an influenza outbreak was present. We first ran BNetScan-PC for every day in the 1-year background period, computed $P(E = influenza|Data)$ on each of those days, where *Data* depends on the day, and then ordered those posterior probabilities in decreasing order. Let $threshold_0, threshold_1, \ldots, threshold_{365}$ be that ordered list.

For a given outbreak, let $Data_i$ be the data obtained on the ith day of the outbreak. For a given false positive rate r, let i_r be the smallest value of i such that

$$P(E = influenza|Data_i) > threshold_r.$$

Then for that outbreak, the days to detection was set equal to $i_r - 1$ at a false alarm rate of r. To obtain the mean days to detection at false alarm rate of r, we took the average of the days-to-detection over all the simulated outbreaks. In order to plot the system's ability to detect an outbreak in general, we follow the same procedure except we use $P(E \neq none|Data_i)$ instead of $P(E = influenza|Data_i)$.

To create the AMOC curves for SaTScan™, the likelihood ratio (Eq. (1)) of the most likely subregion was used instead of a posterior probability.

Figs. 4 and 5 show AMOC curves comparing the performance of the five methods. We can see from these AMOC curves that BNetScan-PC, BSS Method 2, and SaTScan™ Method 2 all typically performed much better than BSS Method 1 and SaTScan™ Method 1. These results support the usefulness of modeling probabilistic relationships in performing outbreak detection. Further, the AMOC curves show that in the case of *Cryptosporidium* outbreaks, BNetScan-PC has better performance than the other methods, and this is particularly true when the number of false alarms per year is fewer than 3. Indeed, BNetScan-PC's performance hardly degrades as we approach a false alarm rate of 0. In the case of influenza outbreaks, the performance of SaTScan™ Method 2 exceeds that of BNetScan-PC, but not by as much as BNetScan-PC's performance exceeds that of SaTScan™ Method 2 in the case of *Cryptosporidium* outbreaks.

Another notable result is that the performance of each of the methods does not degrade very much as the probability distribution used to generate the data increasingly deviates from the distribution defined by the BNetScan-PC model. As expected, the performances of BSS Method 1 and SaTScan™ Method 1 would not degrade, since these methods do not use probabilistic information from BNetScan-PC. However, even when $N = 1$ the mean day at which BNetScan-PC detects a *Cryptosporidium* outbreak is 4.63 when the annual false positive rate is 0, whereas for $N = \infty$ it is 3.56. The corresponding values for influenza outbreak are 10.81 and 8.11, respectively. These results are encouraging, and are consistent with the findings in [18], which indicated that diagnosis using Bayesian networks is often insensitive to imprecision in probabilities. In the case of outbreak detection, we suspect that, as long as we identify some of the most likely chief complaints given an outbreak disease, we can obtain good detection performance even if imprecision is high.

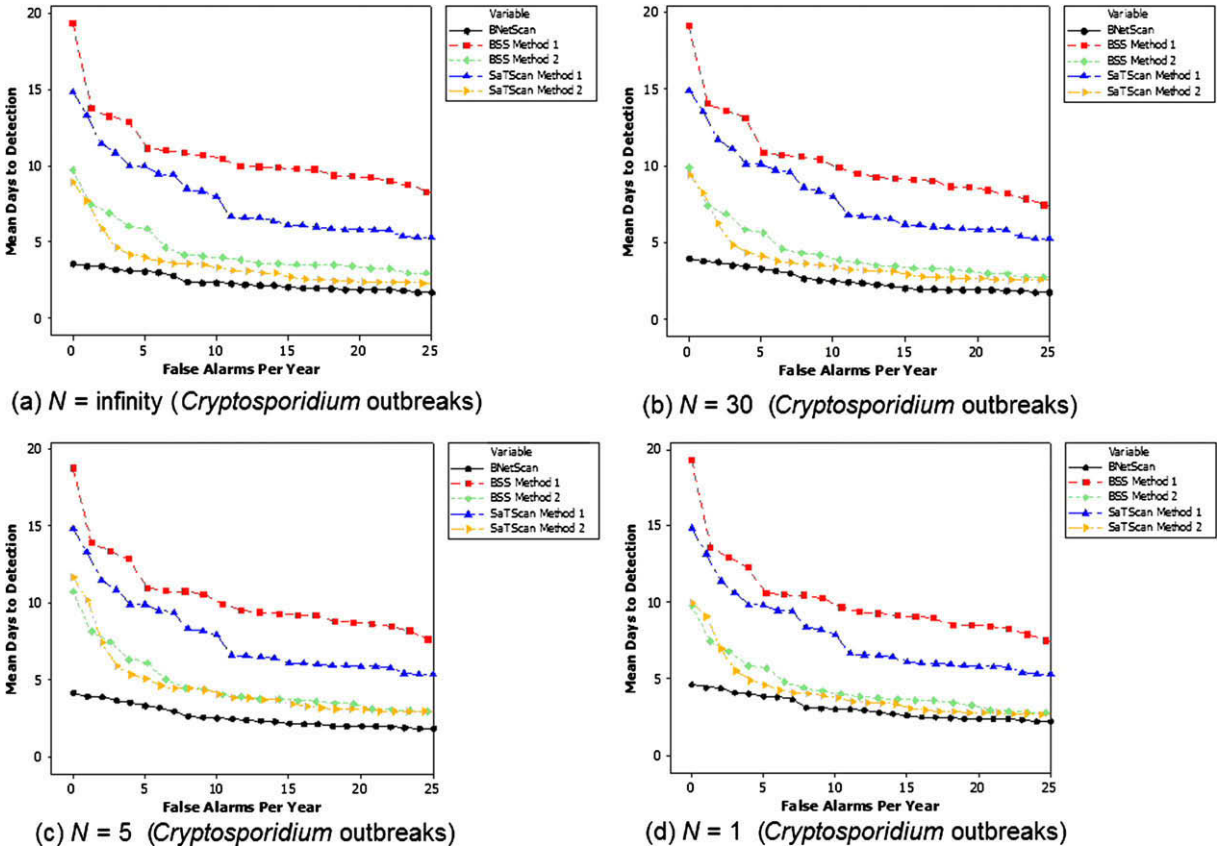

Fig. 4. AMOC curves for *Cryptosporidium* outbreaks.

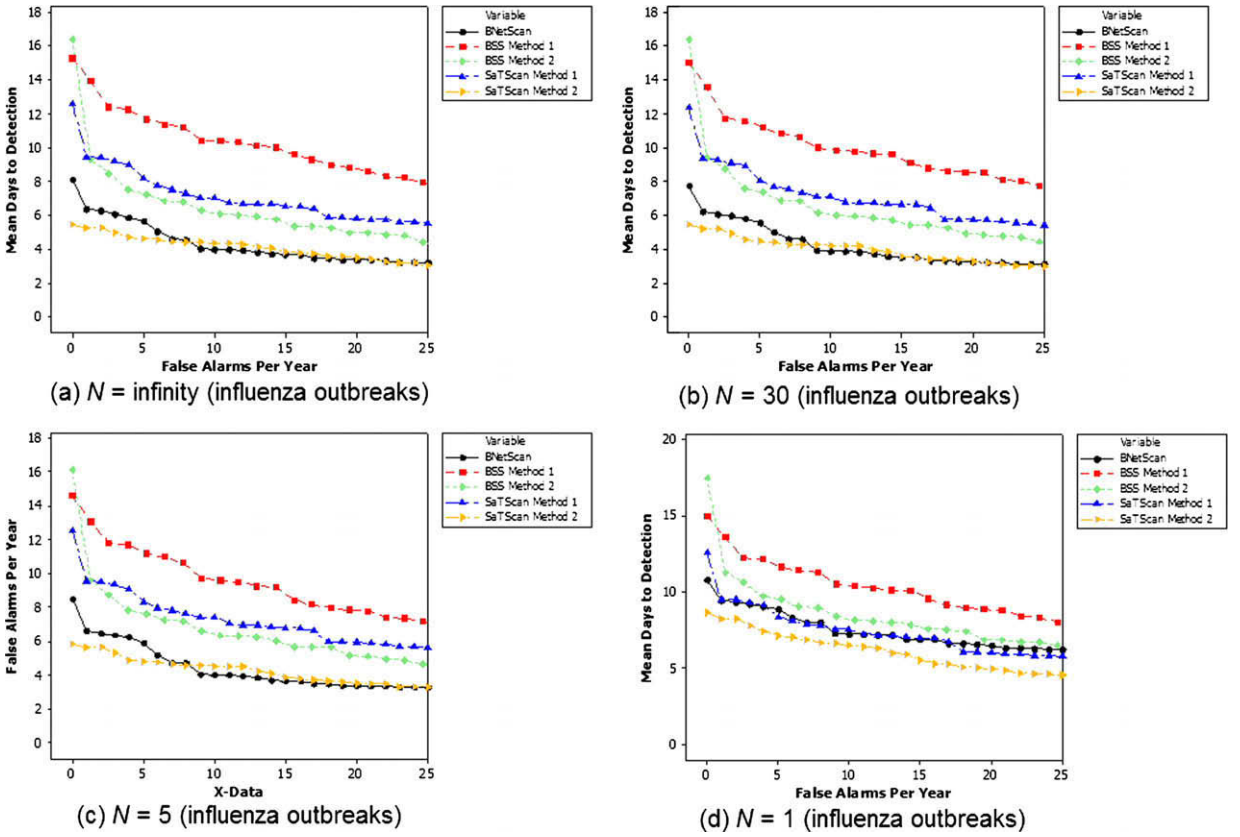

Fig. 5. AMOC curves for influenza outbreaks.

Recall that BNetScan-PC is designed to detect 12 different types of outbreaks. So it not only reports the posterior probability of an influenza or a *Cryptosporidium* outbreak, but also the overall probability of any outbreak being present. Fig. 6 contains AMOC curves showing BNetScan-PC's performance regarding the detection of any outbreak. In that figure, we compare general outbreak detection performance to performance in detecting the disease-specific outbreak. For example, we used $P(E = influenza|Data)$ to produce the AMOC curve in Fig. 6c, and we used $P(E \neq none|Data)$ to produce the AMOC curve in Fig. 6d. In the case of *Cryptosporidium* outbreaks, the system can detect any outbreak almost as well as it can specifically detect a *Cryptosporidium* outbreak. In the case of influenza outbreaks, for most values of N the performance when detecting any outbreak is about the same as that when detecting an influenza outbreak, and when $N = 1$ BNetScan-PC actually detects any outbreak much better than it specifically detects an influenza outbreak. The result for $N = 5$, 30, and ∞ may be due to the fact that influenza outbreaks are be fairly difficult to detect. They are difficult to detect because influenza has symptoms such as cough and fever/chills, which are not uncommon when no outbreak is occurring. Since influenza outbreaks are difficult to detect, it seems reasonable that it would be no easier to detect an influenza outbreak than a non-specific outbreak. The result for $N = 1$ seems reasonable for the same reason. That is, perhaps, when the conditional probability distributions of the chief complaints given influenza in the detection model are significantly different from the conditional probability distributions used to generate the outbreak data, it may be quite difficult to detect an influenza outbreak. However, since there is still a substantial increase in ED visits during the outbreak, it may not be that difficult to detect a non-specific outbreak. Note that *Cryptosporidium*, on the other hand, has very distinct symptoms.

Notice from Figs. 4 and 5 that SaTScan™ Method 2 tended to outperform BSS Method 2. This may be due to the following factor. SaTScan™ searches over circular regions, while BSS searches over rectangles. This difference gives SaTScan™ an advantage in these particular experiments because all of the outbreak subregions were highly compact (near-circular). However, BSS may outperform SaTScan™ if the subregions were highly elongated, which we did not test here.

4.2.2. Subregion detection

We considered the subregion s_j that maximized $P(Data|S = s_j)$ to be the subregion detected by BNetScan-PC and BSS. Note that this is the same subregion with maximum posterior probability if we assume that all subregions have the same prior probability. We considered the subregion that maximized the Poisson spatial scan statistic to be the subregion detected by SaTScan™. To measure the accuracy of the detections, we used the *overlap coefficient* of the detected subregion with the injected (actual) subregion. That overlap coefficient is defined as follows:

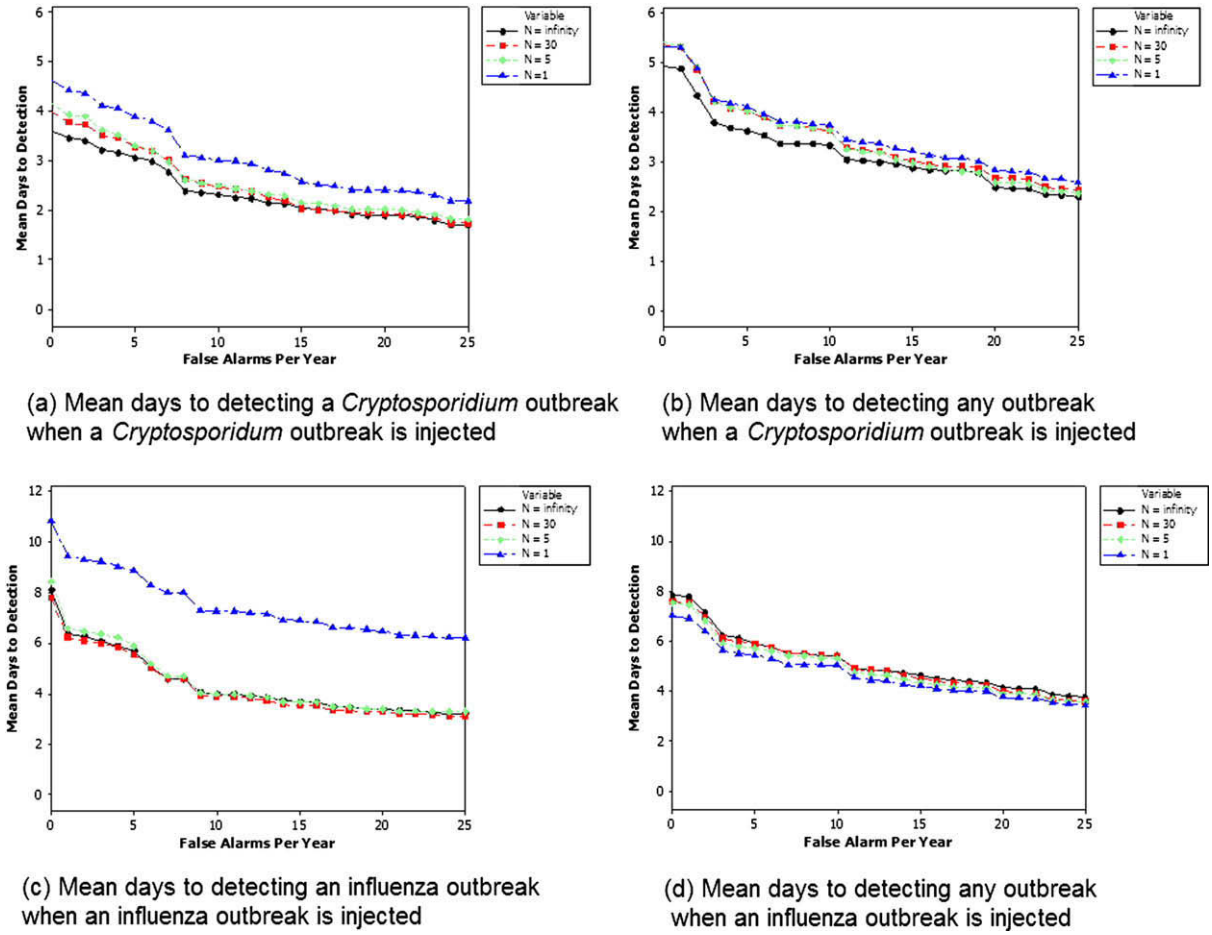

Fig. 6. AMOC curves comparing BNetScan-PC's ability to detect any outbreak (non-disease specific) to its ability to detect the specifically simulated outbreak disease.

$$overlap(s_1, s_2) = \frac{\#(s_1 \cap s_2)}{\#(s_1 \cup s_2)},$$

where # returns the number of cells in a region. This coefficient is 0 if and only if the two subregions do not intersect, while it is 1 if and only if they are the same subregion.

Figs. 7 and 8 show the average values of the overlap coefficients on each day of the outbreaks. In all cases the performance of BNetScan-PC was significantly better than of the other methods. In the case of *Cryptosporidium* outbreaks, BNetScan-PC performed substantially better than SaTScan™ Method 2 at both outbreak detection and subregion detection. However, in the case of influenza outbreaks, BNetScan-PC performed worse than SaTScan™ Method 2 at outbreak detection, but it substantially outperformed SaTScan™ Method 2 at subregion detection. One reason for this enigmatic result can be understood by looking at Fig. 6c and d. BNetScan-PC was substantially better at detecting any outbreak than it was at specifically detecting an influenza outbreak. SaTScan™ Method 2 does not report which type of outbreak it detects; it only looks for a cluster of the top three chief complaints. If we had compared BNetScan-PC's ability to detect any outbreak to the performance of SaTScan™ Method 2, BNetScan-PC would have performed better than SaTScan™ Method 2 at outbreak detection. This is consistent with its superior subregion detection performance, which is not sensitive to which particular disease is driving the high probability of an outbreak.

BNetScan-PC's performance did not degrade as the probability distribution used to generate the data increasingly deviated from the one in the BNetScan-PC model (i.e., as N went from ∞ to 1). These results are understandable in that the number of cases injected into the outbreak subregion does not depend on N. That is, if we changed the type of chief complaints injected but not their number, we may detect a different outbreak, but we would not change the subregion most likely to contain the outbreak. In the same way, the performance of Methods 1 of SaTScan™ and BSS should not deteriorate as we decrease N. However, Figs. 7 and 8 show that the performance of the Method 2 versions of BSS and SaTScan do deteriorate somewhat, which can be explained by the fact that they only consider a subset of chief complaints that are consistent with

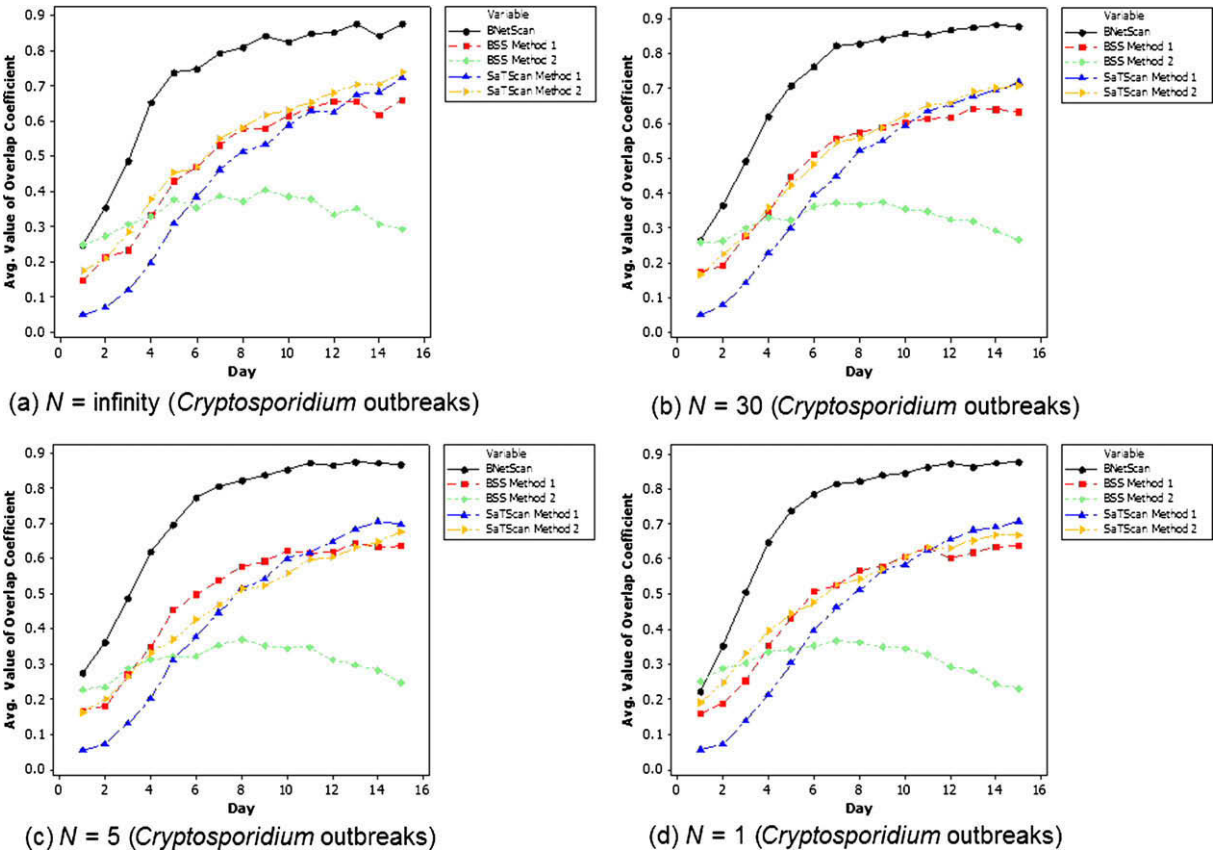

Fig. 7. The average value of the overlap coefficient for *Cryptosporidium* outbreaks.

the simulated outbreak disease; thus, that set may become less informative about the presence of an outbreak as N decreases from ∞ to 1.

Another noteworthy result is that the spatial-localization performance of Method 1 for SaTScan™ and BSS is sometimes better than the performance of Method 2. This is particularly true when the variable *Day* is large, which means the number of outbreak cases becomes large. The explanation for this result can be appreciated by considering SaTScan™. When we are only determining the most likely subregion, we are comparing each subregion to each other subregion. We are not concerned with how the likelihood of the most likely subregion compares to the likelihoods obtained during the background period. Consider the Poisson spatial scan statistic (Eq. (1)). Method 1 counts more chief complaints than Method 2, and therefore in the outbreak subregion S the value of $C_{in}^{(s)}$ will be larger for Method 1 than for Method 2. This should often increase the value of the statistic more for the outbreak subregion S than for the other subregions, especially when the additional counts due to the additional chief complaints is large. On the other hand, when we are merely detecting an outbreak (and not characterizing its location), we are concerned with how the likelihood compares to likelihoods computed during the background period.

5. Summary

We compared the performance of three spatial cluster detection methods in the context of disease-outbreak surveillance. They are a frequentist spatial scan statistic (we used the SaTScan™ software implementation of this method), a Bayesian spatial scan statistic (BSS), and a novel Bayesian-network-based spatial scan statistic (BNetScan), which was introduced in this paper. In the majority of cases, BNetScan outperformed the other two methods both in terms of outbreak detection and subregion detection. This was the case even when SaTScan™ and BSS were able to take advantage of the probabilistic information in BNetScan. These results lend support to the conjecture that, in the case of spatial event surveillance, we may be able to obtain better results by modeling the relationships among the events of interest and the observable events using a Bayesian network rather than using a summary statistic.

We also found that the performance of BNetScan was robust relative to the probability distribution generating the data. Thus, the results support that in the domain of disease-outbreak detection we may expect to be able to achieve good detection performance even when the assumed and actual probability distributions are quite different.

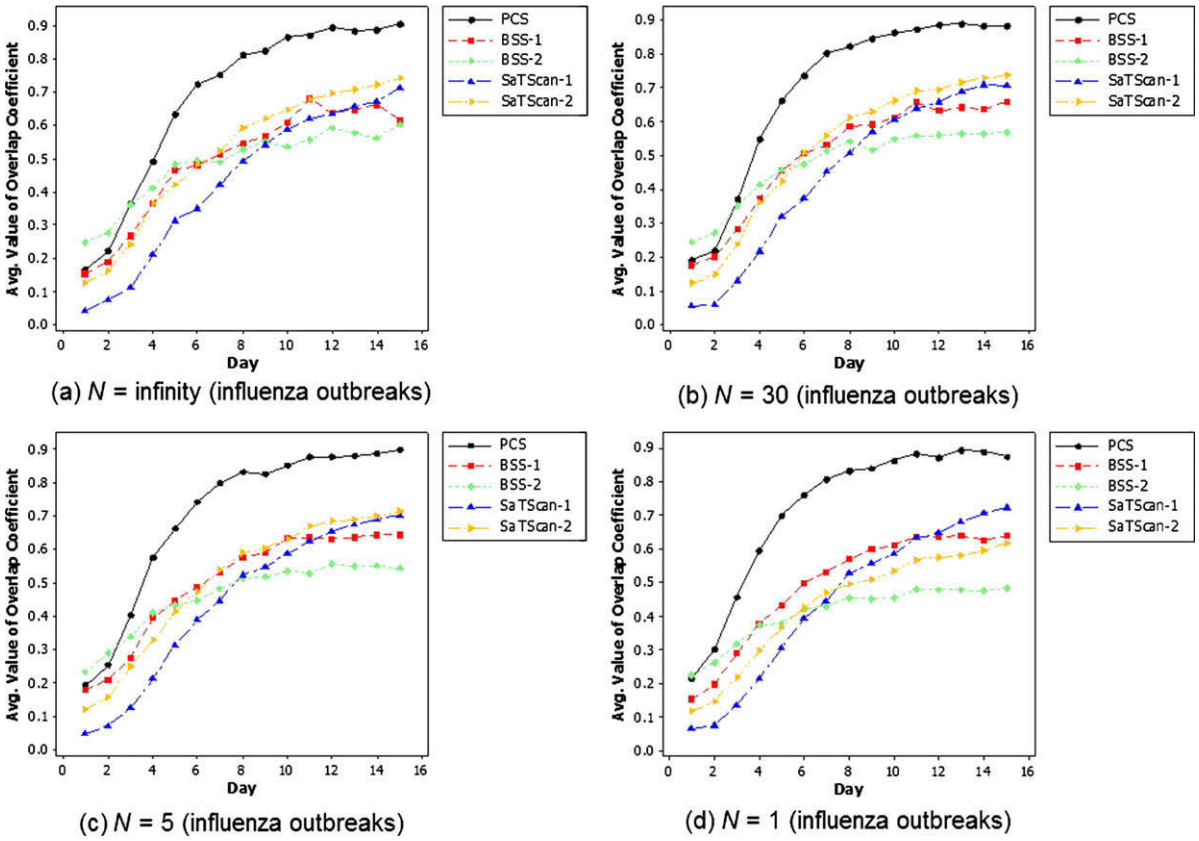

Fig. 8. The average value of the overlap coefficient for influenza outbreaks.

Finally, we found that BNetScan can detect the presence of any outbreak almost as well as (indeed, sometimes better) it detects a specific outbreak. It will be useful to investigate this matter further by performing experiments in which all 12 types of outbreaks represented in BNetScan-PC are simulated.

Acknowledgements

This research was supported by the National Science Foundation Grant No. IIS-0325581.

References

[1] M. Kulldorff, A spatial scan statistic, Communications in Statistics: Theory and Methods 26 (6) (1997) 1481–1496.
[2] M. Kulldorff, Satscan v. 4.0: Software for the Spatial and Space–time Scan Statistics, Technical Report, Information Management Services Inc., 2004.
[3] D.B. Neill, A.W. Moore, G.F. Cooper, A Bayesian spatial scan statistic, Advances in Neural Information Processing Systems (NIPS) 18 (2006) 1003–1010.
[4] M. Kulldorff, N. Nagarwalla, Spatial disease clusters: detection and inference, Statistics in Medicine 14 (1995) 799–810.
[5] J.I. Naus, The distribution of the size of the maximum cluster of points on the line, Journal of the American Statistical Association 60 (1965) 532–538.
[6] M. Kulldorff, E.J. Feuer, B.A. Miller, L.S. Freedman, Breast cancer clusters in the Northeast United States: a geographical analysis, American Journal of Epidemiology 146 (2) (1997) 161–170.
[7] U. Hjalmars, M. Kulldorff, G. Gustafsson, N. Nagarwalla, Childhood leukemia in Sweden using GIS and a spatial scan statistic for cluster detection, Statistics in Medicine 15 (1996) 707–715.
[8] M. Kulldorff, Z. Fang, S.J. Walsh, A tree-based scan statistic for database disease surveillance, Biometrics 59 (2003) 323–331.
[9] F. Mostashari, M. Kulldorff, J.J. Hartman, J.R. Miller, V. Kulasekera, Dead bird clustering: a potential early warning system for west nile virus activity, Emerging Infectious Diseases 9 (2003) 641–646.
[10] I. Jung, M. Kulldorff, A. Klassen, A spatial scan statistic for ordinal data, Statistics in Medicine 26 (2006) 1594–1607.
[11] D.B. Neill, A.W. Moore, G.F. Cooper, A multivariate Bayesian spatial scan statistic, Advances in Disease Surveillance 2 (2007) 60.
[12] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunco, F. Mostashari, Space–time permutation scan statistic for disease outbreak detection, PLoS Medicine 2 (2005) 216–224.
[13] D.B. Neill, A.W. Moore, M. Sabnani, K. Daniel, Detection of emerging space–time clusters, in: Proceedings of 11th ACM SIGKDD International Conference on Knowledge Discovery and Mining, Chicago, IL, 2005, pp. 218–227.
[14] M. Dwass, Modified randomization tests for nonparametric hypotheses, Annals of Mathematical Statistics 28 (1957) 181–187.
[15] B.W. Turnball, E.J. Iwano, W.S. Burnett, H.L. Howe, L.C. Clark, Monitoring for clusters of disease: application to leukemia incidence in upstate New York, American Journal of Epidemiology 132 (1990) 136–143.
[16] G.F. Cooper, J.N. Dowling, J.D. Lavender, P. Sutovsky, A Bayesian Algorithm for Detecting CDC Category A Outbreak Diseases from Emergency Department Chief Complaints, Advances in Disease Surveillance 2 (2007) 45.

[17] T. Fawcett, F. Provost, Activity monitoring: noticing interesting changes in behavior, in: Proceedings of the Fifth SIGKDD Conference on Knowledge Discovery and Data Mining, San Diego, CA, 1999, pp. 53–62.
[18] M. Henrion, M. Pradhan, B. Del Favero, K. Huang, G. Provan, P. O'Rorke, Why is diagnosis using belief networks insensitive to imprecision in probabilities? in: E. Horvitz, F. Jensen (Eds.), Uncertainty in Artificial Intelligence; Proceedings of the Twelfth Conference, Morgan Kaufmann, Burlington, MA, 1996, pp. 307–314.